LLNL ASCI Resources

Blaise Barney

LLNL

Services & Development Division

Los Alamos NATIONAL LABORATORY

Sandia National Laboratories

NNSA National Nuclear Security Administration

Lawrence Livermore National Laboratory

SDSC

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

THE UNIVERSITY OF CHICAGO

NNSA ASCI Advanced Simulation and Computing

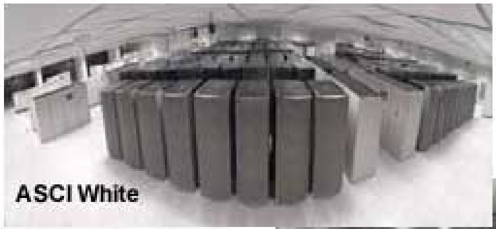
Los Alamos

Sandia

Lawrence Livermore

LELAND STANFORD JUNIOR UNIVERSITY ORGANIZED 1891

THE UNIVERSITY OF UTAH


ASCI White

ASCI Blue-Pacific

ASCI Blue Mountain

ASCI Red

NPACI Blue Horizon

ASCI Cplant

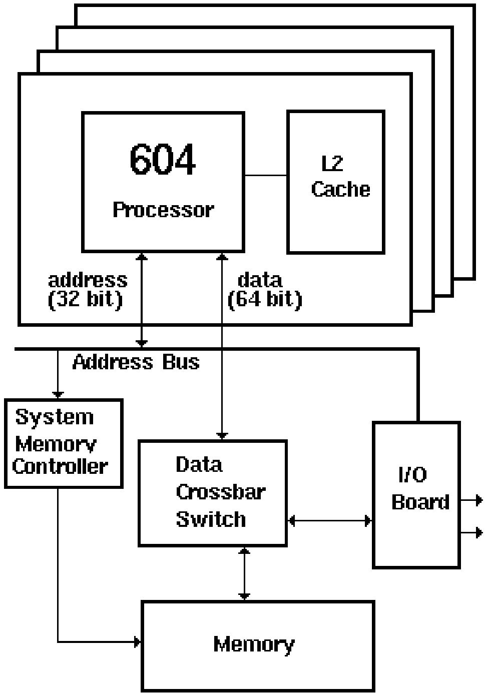
Overview

- **Hardware Environment**
  - ASCI Blue-Pacific
  - ASCI White
  - Berg
  - ASCI Purple
  - ALC
  - Parallel File Systems
  - HPSS
  - Blue/Frost Batch Queues
  - Alliance YTD Usage
- **Software Environment**
  - AIX, PSSP, CHAOS, SLURM
  - Compilers
  - Other Software
- **Training**
- **Futures**
  - Future Plans
  - Blue Gene/L
  - Terascale Simulation Facility

ASCI Blue-Pacific

- **Blue (OCF)**
  - 264 total nodes
  - 256 compute nodes
  - 1.5 GB memory/node
  - 17+ TB parallel file system

- **SKY (SCF)**
  - 3 sectors of 488 nodes - 1464 total nodes
  - 1296 compute nodes
  - 1.5 - 2.5 GB memory/node
  - 62+ TB parallel file system (total)



- **IBM 604e technology**
  - 4 CPUs/node
  - 332 MHz clock
  - 664 Mflops/CPU
  - 256 KB L2 cache/CPU
  - 32-bit architecture

- **Stable production systems**

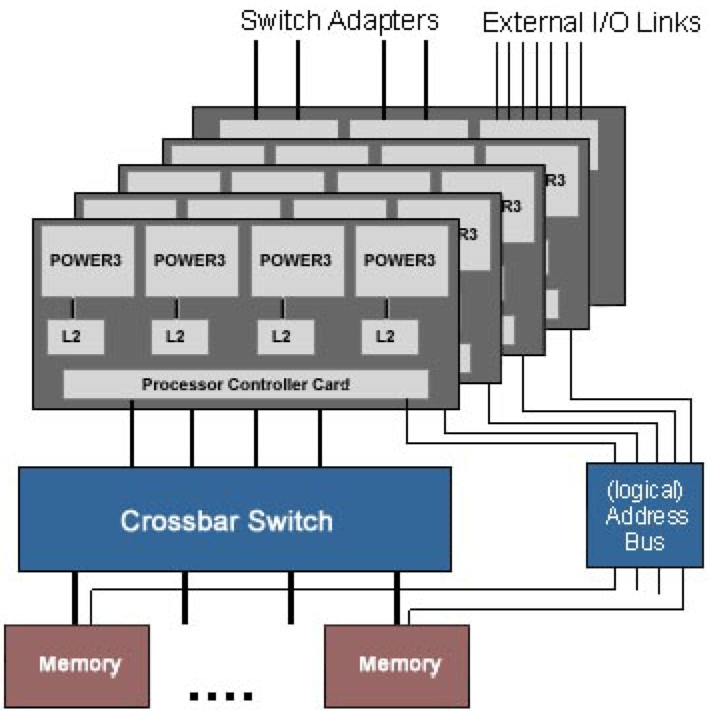
ASCI White

- **Frost (OCF)**
  - 68 total nodes
  - 64 compute nodes
  - 16 GB memory/node
  - 20+ TB parallel file system

- **White / Ice (SCF)**
  - 512 / 28 total nodes
  - 489 / 26 compute nodes
  - 16 GB memory/node
  - 109 / 5.7 TB parallel file systems

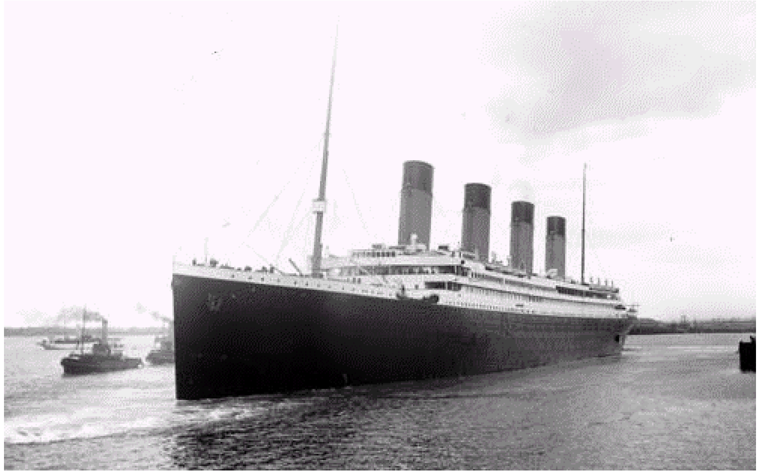
- **IBM POWER3 technology**
  - 16 CPUs/node
  - 375 MHz clock
  - 1500 Mflops/CPU
  - 8MB L2 cache/CPU
  - 64-bit architecture

- **Stable production systems**



Berg

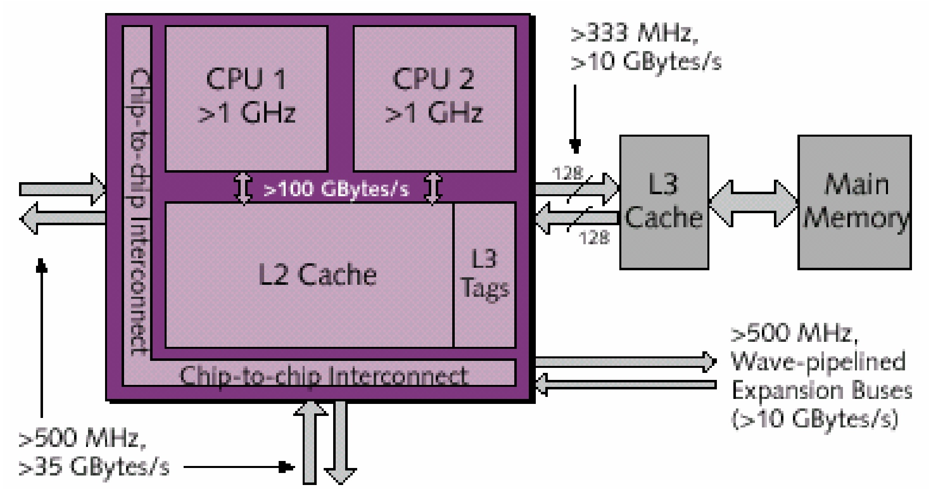
- **OCF: Berg**
  - 2 nodes
  - 32 GB memory/node
- **IBM POWER4 technology**
  - 32 CPUs/node
  - 1.3 GHz clock
  - 5200 Mflops/CPU
  - 23 MB L2 cache/node
  - 512 MB L3 cache/node
  - 64-bit architecture



- **Available**
  - Stable, but not a production system
  - Primarily intended for pre-Purple POWER4 testing
  - Configured into four virtual machines of 14, 16, 2 and 32 nodes
  - No switch or parallel file system
  - Both interactive and batch usage
  - Available for Alliance users

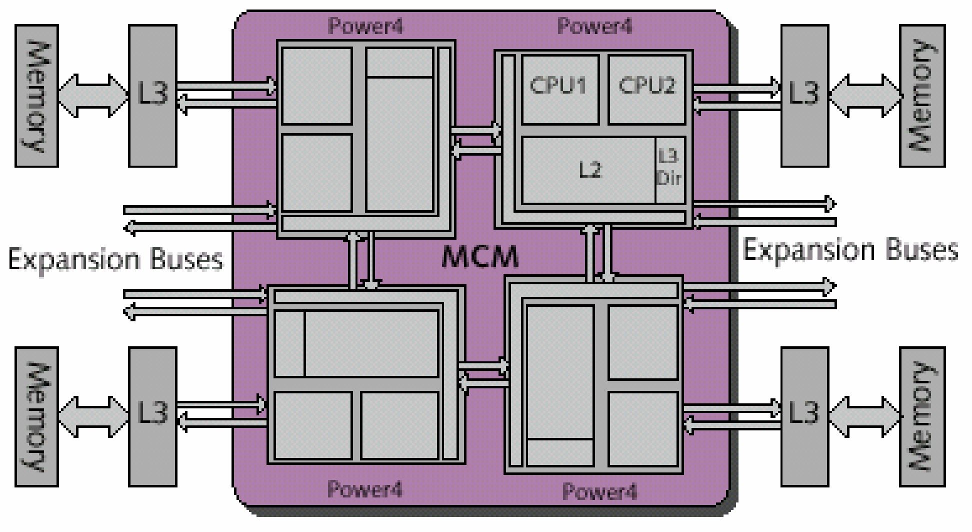
POWER4

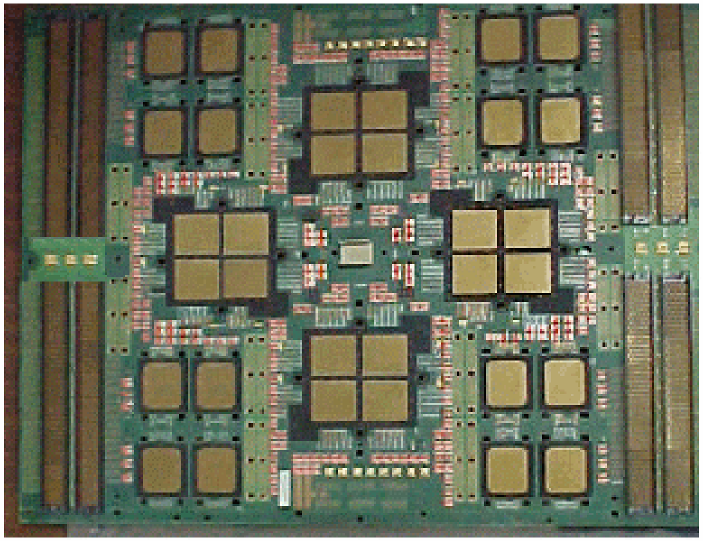
- **2 CPUs per chip**
  > 100 GB/s L2 bandwidth
  > 10 GB/s L3 bandwidth



- **4 chips per module**
  > 35 GB/s chip-to-chip

- **4 modules = 32-way SMP**
  - 20 GB/s module-to-module
  - Logically shared L2 and L3 cache





ASCI Purple

- **ASCI Option Purple**
  - 5th generation ASCI platform
  - 60+ teraOPs system
  - Option for 100+ teraOPs system

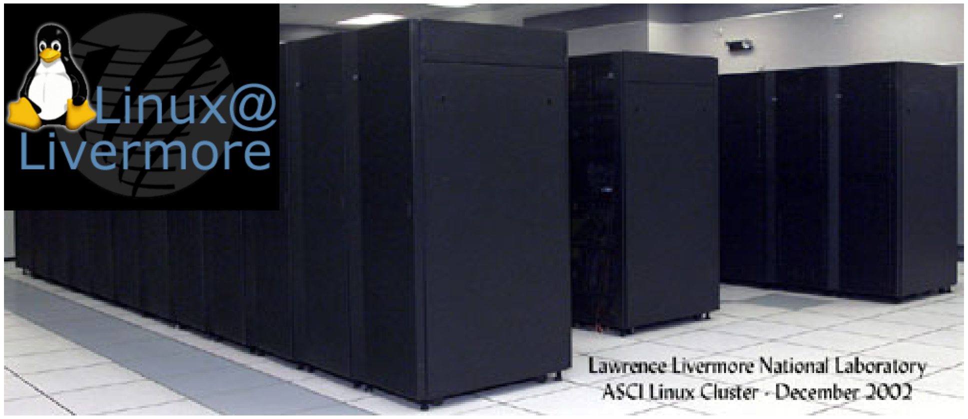
- **Update**
  - 9.2 Tflop unclassified component has been delivered.
  - Classified Ed-tv (Early Delivery Technology Vehicle) component will start arriving by mid-September. Will consist of 256 8-way Power4 nodes by December 2003.
  - Final delivery of ~60-100 teraOPs Power5 system still scheduled for late 2004

- **During Ed-tv's testing phase in the OCF, Alliances can request to participate in "science" runs phase**

- **SCF system will reside in the new Terascale Simulation Facility**

ALC +?

- **ASCI Linux Cluster**
  - Unclassified component of ASCI Purple
  - 9.2 TFlop system
  - 960 nodes
  - Each node has 2 Pentium4 Xeon (Prestonia) processors
  - 4 GB memory per node
  - Quadrics switch

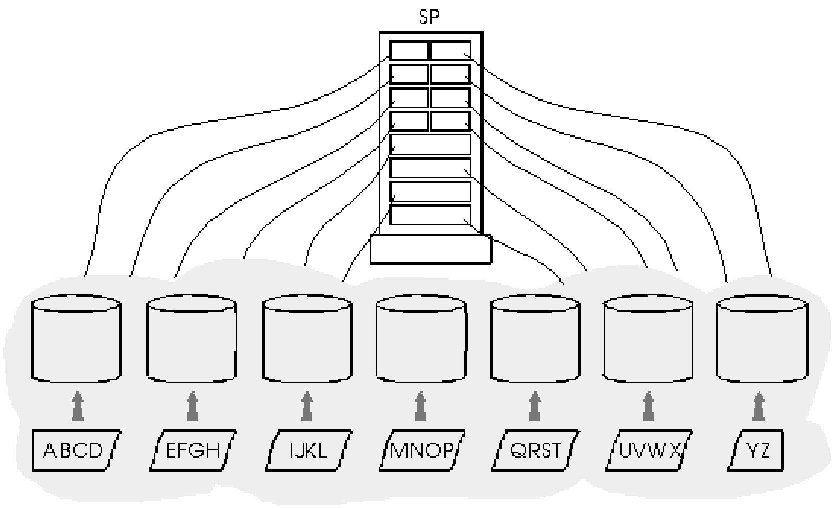

Linux@
Livermore
Lawrence Livermore National Laboratory
ASCI Linux Cluster - December 2002

- **Schedule**
  - Operational since Feb 2003
  - Currently in testing and development phase – primarily for the Luster parallel file system
  - Limited availability (LA) phase may begin in October

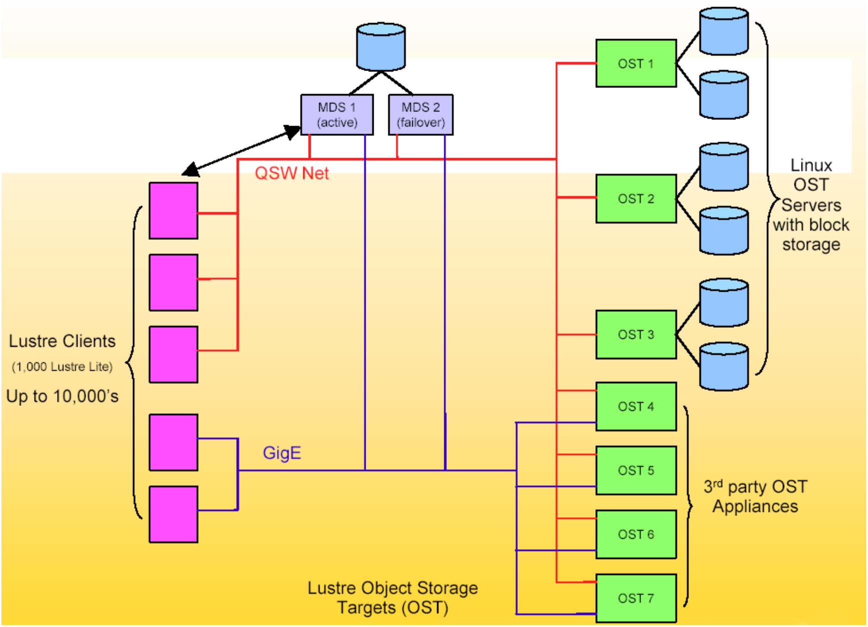
- **Alliance Resource**
  - Some Alliance codes may be candidates for LA phase
  - 5 -10% of ASCI Purple Tflops will be made available to Alliances on ALC and other to be defined resources
  - Blue/Frost users will automatically have an account on ALC when GA
  - Queues and limits TBD

Parallel File Systems

- **GPFS**
  - IBM's General Parallel File System
  - All LLNL ASCI systems have their own, multi-terabyte GPFS file system(s)
  - Frost performance with 60 client nodes and 2 server nodes:
    Write = 550 MB/s
    Read = 600 MB/s

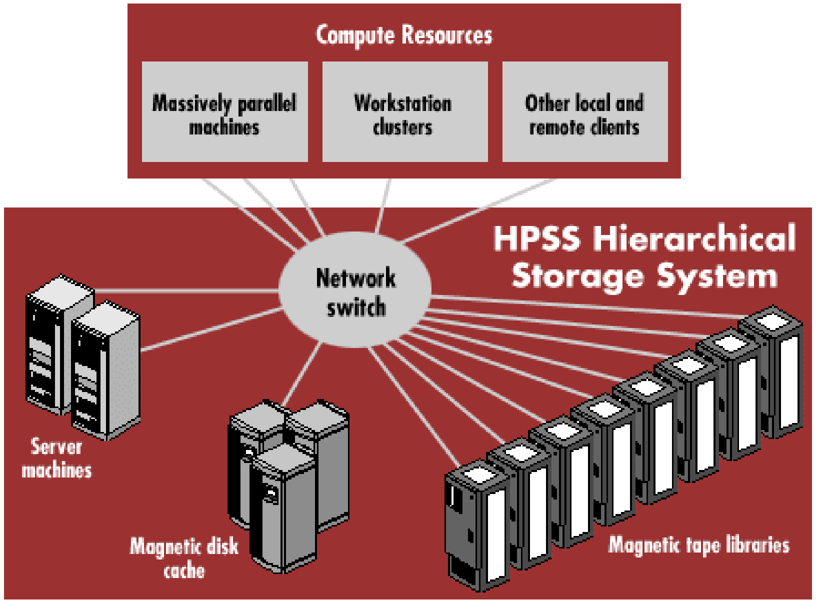


- **Lustre**
  - Linux cluster based parallel file system from Cluster File Systems, Inc.
  - Goals: clusters with 10,000s of nodes, petabytes of storage, move 100s of GB/s with state of the art security and management infrastructure.
  - Currently running on Livermore's ALC and MCR machines.
  - Still being developed and tested
  - See **www.lustre.org** for more info

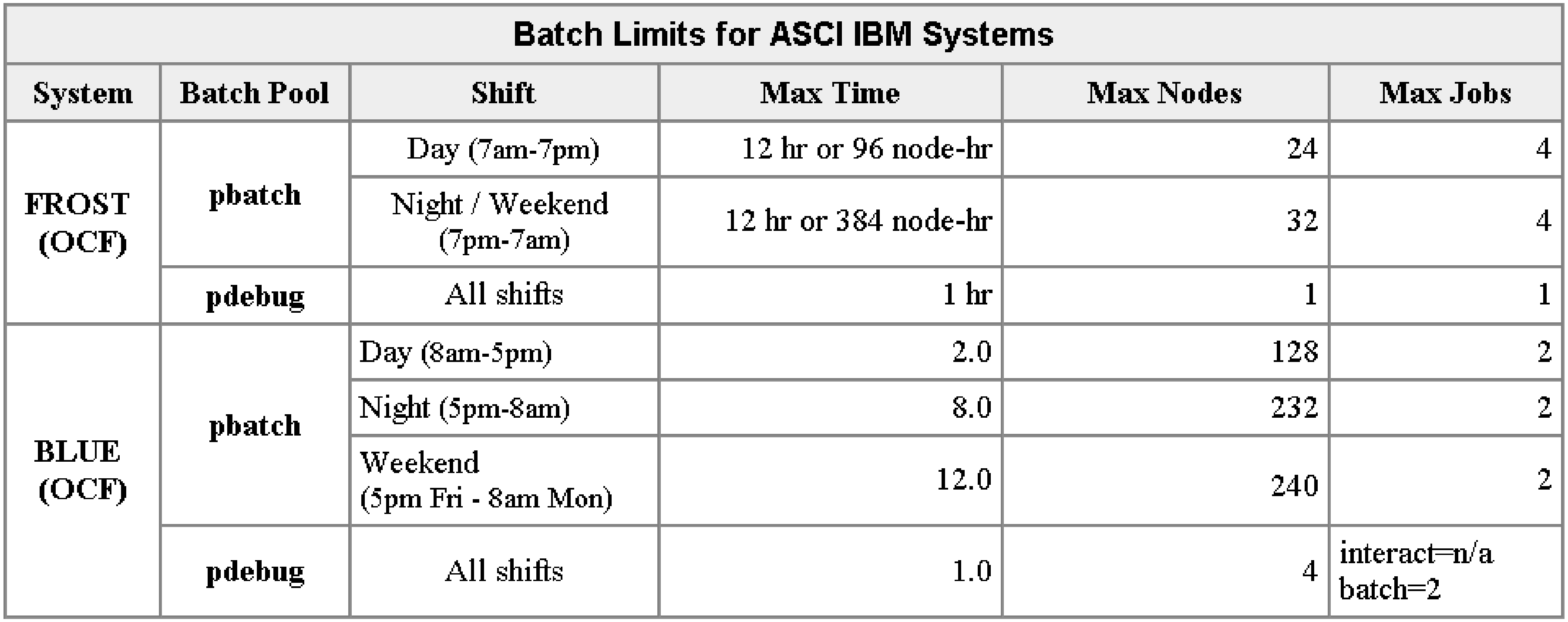


HPSS Archival Storage

- **Integrated into the OCF and SCF gigabit ethernet networks**
- **Some metrics**
  - OCF: 1.8 PB capacity
    @431 TB used
  - SCF: 2.7 PB capacity
    @924 TB used
  - 250 MB/s aggregate writes to HPSS
  - 150 MB/s on a per file basis



Blue and Frost Batch Queues



Batch Limits for ASCI IBM Systems

System	Batch Pool	Shift	Max Time	Max Nodes	Max Jobs
FROST (OCF)	pbatch	Day (7am-7pm)	12 hr or 96 node-hr	24	4
		Night / Weekend (7pm-7am)	12 hr or 384 node-hr	32	4
	pdebug	All shifts	1 hr	1	1
BLUE (OCF)	pbatch	Day (8am-5pm)	2.0	128	2
		Night (5pm-8am)	8.0	232	2
		Weekend (5pm Fri - 8am Mon)	12.0	240	2
	pdebug	All shifts	1.0	4	interact=n/a batch=2

ASCI Alliance YTD Usage

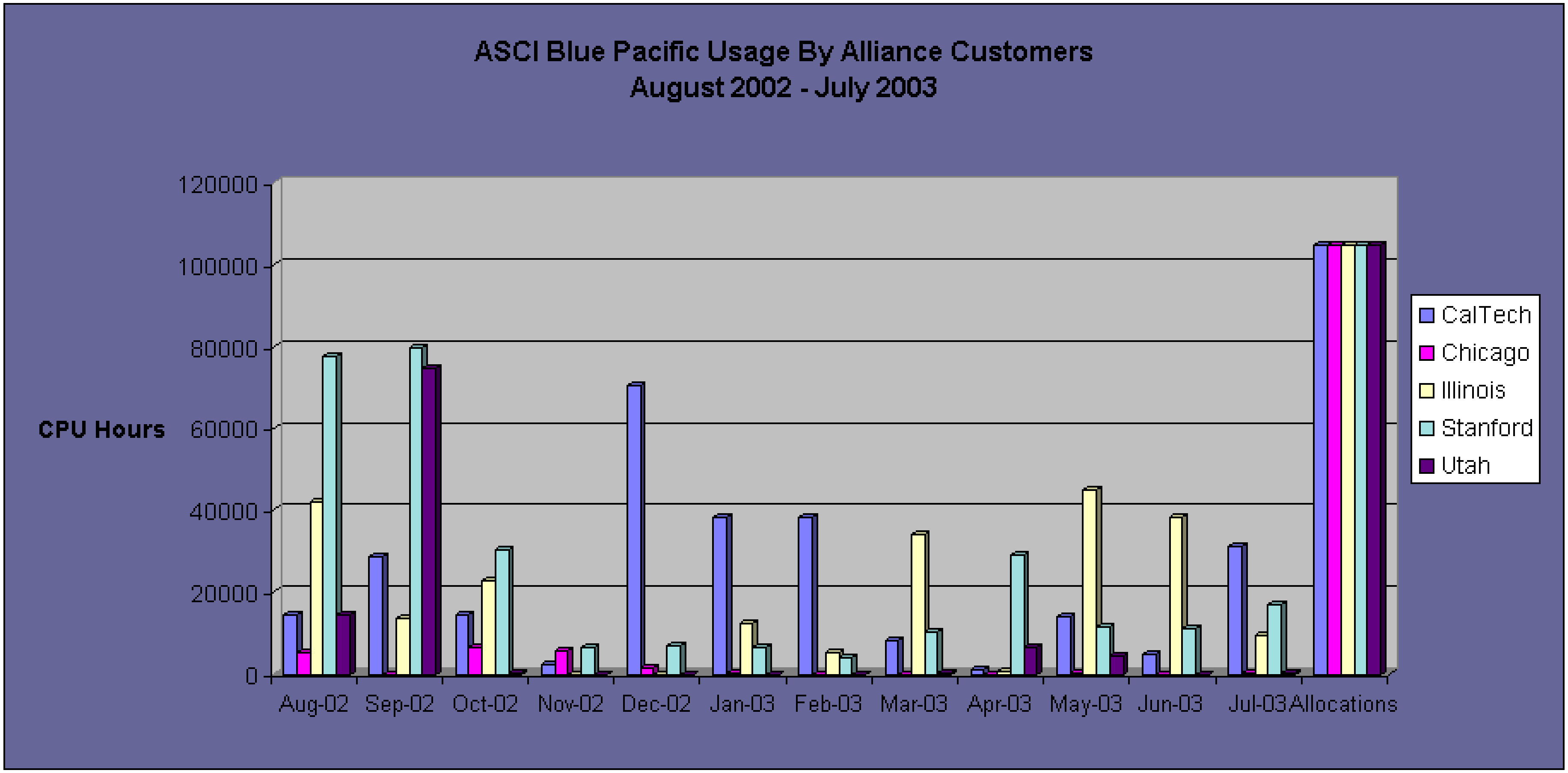
- **Allocations**

  Blue: 104,725 hr/month per alliance
  Frost: 57,600 hr/month per alliance

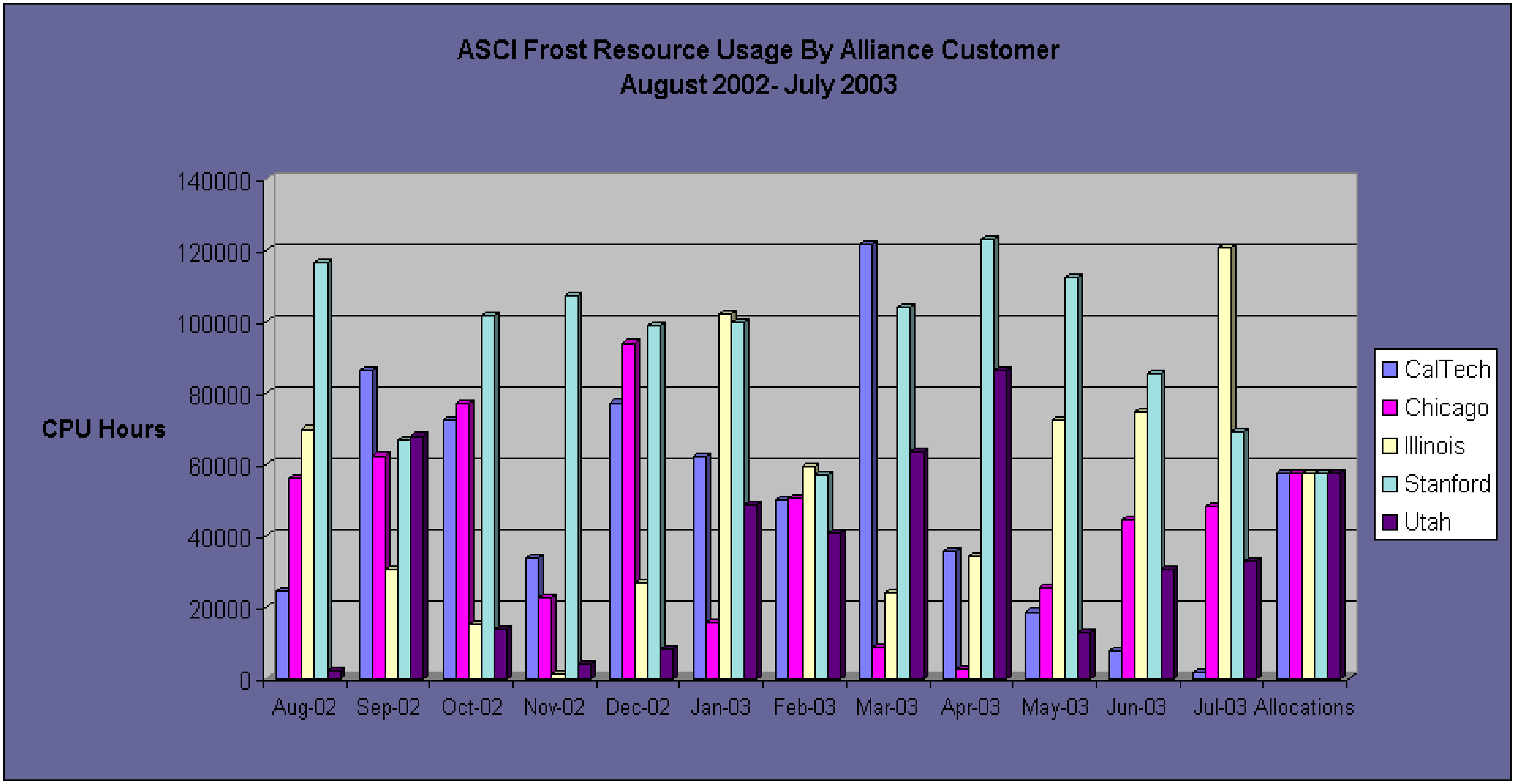
- **Usage 8/02 – 7/03**

Alliance	**Blue**		**Frost**	
	hrs	**%**	**hrs**	**%**
Caltech	265740	21	590835	85
Chicago	19246	2	507156	73
Illinois	223093	18	631010	91
Stanford	290780	23	1140483	165
Utah	100485	8	412121	60
Total	899344	14	3281605	95

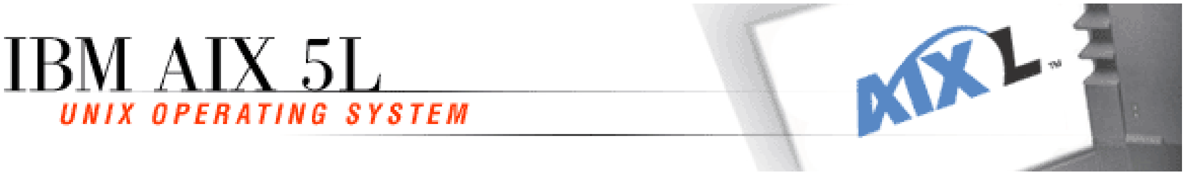
ASCI Alliance YTD Usage



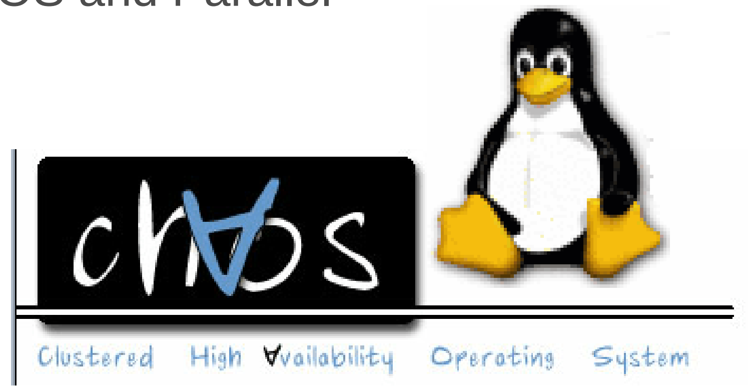
ASCI Alliance YTD Usage



Software Environment



- **AIX 5.1   PSSP 3.4**
  - All ASCI IBM systems are now at the same OS and Parallel Environment software levels

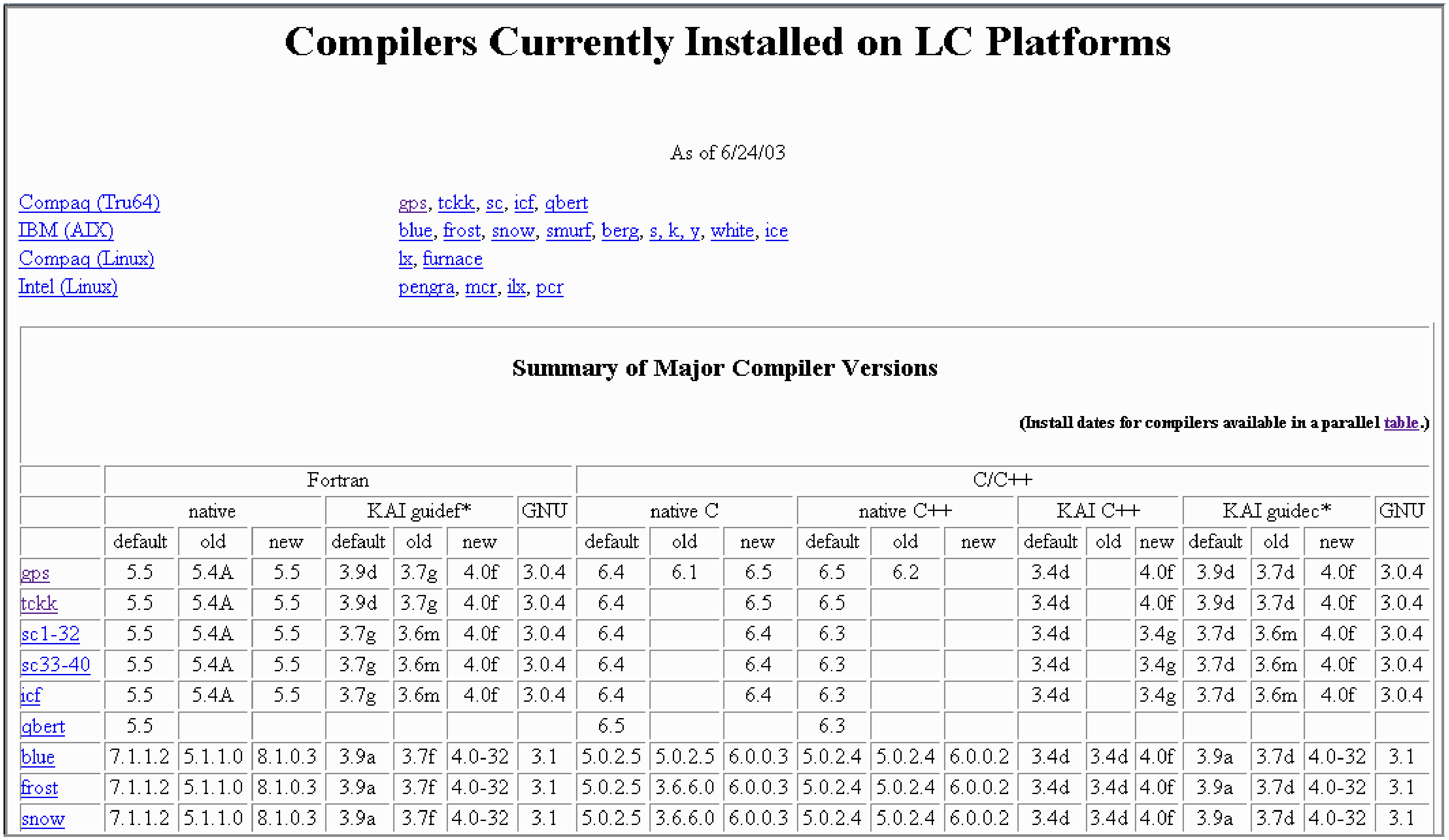


- **CHAOS**
  - Clustered High Availability Operating System
  - LC's developmental Linux cluster OS
  - Based upon RedHat Linux (currently 7.3)
  - Used on ALC and all other LC Linux clusters

- **SLURM**
  - Simple Linux Utility for Resource Management
  - Collaboration between Livermore and Linux NetworX
  - Under development
  - Used on ALC and all other LC Linux clusters

Software Environment

- **Compilers**
  - Fortran
  - C / C++
  - With MPI, OpenMP



Compilers Currently Installed on LC Platforms

As of 6/24/03

Compaq (Tru64)	gps, tckk, sc, icf, qbert
IBM (AIX)	blue, frost, snow, smurf, berg, s, k, y, white, ice
Compaq (Linux)	lx, furnace
Intel (Linux)	pengra, mcr, ilx, pcr

Summary of Major Compiler Versions

(Install dates for compilers available in a parallel table.)

	Fortran							C/C++												
	native			KAI guidef*			GNU	native C			native C++			KAI C++			KAI guidec*			GNU
	default	old	new	default	old	new		default	old	new	default	old	new	default	old	new	default	old	new	
gps	5.5	5.4A	5.5	3.9d	3.7g	4.0f	3.0.4	6.4	6.1	6.5	6.5	6.2		3.4d		4.0f	3.9d	3.7d	4.0f	3.0.4
tckk	5.5	5.4A	5.5	3.9d	3.7g	4.0f	3.0.4	6.4		6.5	6.5			3.4d		4.0f	3.9d	3.7d	4.0f	3.0.4
sc1-32	5.5	5.4A	5.5	3.7g	3.6m	4.0f	3.0.4	6.4		6.4	6.3			3.4d		3.4g	3.7d	3.6m	4.0f	3.0.4
sc33-40	5.5	5.4A	5.5	3.7g	3.6m	4.0f	3.0.4	6.4		6.4	6.3			3.4d		3.4g	3.7d	3.6m	4.0f	3.0.4
icf	5.5	5.4A	5.5	3.7g	3.6m	4.0f	3.0.4	6.4		6.4	6.3			3.4d		3.4g	3.7d	3.6m	4.0f	3.0.4
qbert	5.5							6.5			6.3									
blue	7.1.1.2	5.1.1.0	8.1.0.3	3.9a	3.7f	4.0-32	3.1	5.0.2.5	5.0.2.5	6.0.0.3	5.0.2.4	5.0.2.4	6.0.0.2	3.4d	3.4d	4.0f	3.9a	3.7d	4.0-32	3.1
frost	7.1.1.2	5.1.1.0	8.1.0.3	3.9a	3.7f	4.0-32	3.1	5.0.2.5	3.6.6.0	6.0.0.3	5.0.2.4	5.0.2.4	6.0.0.2	3.4d	3.4d	4.0f	3.9a	3.7d	4.0-32	3.1
snow	7.1.1.2	5.1.1.0	8.1.0.3	3.9a	3.7f	4.0-32	3.1	5.0.2.5	3.6.6.0	6.0.0.3	5.0.2.4	5.0.2.4	6.0.0.2	3.4d	3.4d	4.0f	3.9a	3.7d	4.0-32	3.1

Other Software

- **Debuggers, correctness tools**

  - Assure - décor - Great Circle - Insure++
  - TotalView - Umpire -pdbx - ZeroFault

- **Mathematical**

  - IBM optimized libraries: BLAS, ESSL, PESSL
  - Intel optimized MKL on Linux clusters
  - Non-commercial libraries: FFTW, MSSL, MSSL3, PMATH,
  - Packages: LAPACK, METIS, ParMETIS, PETSc
  - LINMATH: Livermore Interactive Numerical Mathematical Software Access Utility

- **Performance analysis**

  - DPCL - Dimemas - Paradyn - Paraver
  - PE Benchmarker - Tau - Vampir/Guideview

- **Profiling**

  - gprof - HPM - MPX - mpiP
  - papi - prof -Xprofiler

- **Visualization, graphics, more...**

Training

- **Regular introductory workshops at LLNL**
  - Parallel programming
  - Linux & Compaq clusters
  - POE
  - Pthreads
  - TotalView
  - LC resources and environment
  - IBM hardware/software
  - MPI
  - OpenMP
  - …

- **Other workshops**
  - Performance analysis tools and topics for the IBM SP
  - MPI performance topics
  - Vampir/GuideView, Paraver, Dimemas
  - Advanced topics for ASCI White users
  - Advanced TotalView
  - Python, Linux topics (planned)

Using LC's ASCI IBM SP and Compaq/Linux Cluster Systems

COMPAQ
IBM
intel
ASCI White
Lawrence Livermore National Laboratory
Lawrence Livermore National Laboratory
MCR Cluster · October 2002

- **Tri-lab and Alliance workshops**
  - Combined training for multiple ASCI platforms held at any Tri-lab or Alliance location
  - Customized workshops delivered at Alliance’s location

Future Plans

- **Blue**
  - Hardware maintenance discontinued. Now using a frame of nodes as spare parts.
  - No date currently set to decommission

- **Frost**
  - Will continue to upgrade AIX and PSSP as long as possible
  - No plans to decommission any time in the foreseeable future

- **Purple**
  - ALC and PVC (viz. cluster) going LA @10/1.
  - Arrival of Ed-tv machines (violet and magenta) mid 9/03

- **Training**
  - Access Grid node to be set up in LC's training center

Blue Gene/L

- **Blue Gene/L**
  - BlueGene/L is a computational sciences research and evaluation platform designed by IBM research for the DOE/NNSA ASCI Program
  - New architecture optimized for cost, performance and scalability
  - 180-360 Tflops
  - 65,536 dual processor nodes with 512 MB memory/node; Torus network
  - IBM PowerPC ASIC processor @700MHz; dual FPU
  - More info: **www.llnl.gov/asci/platforms/bluegenel**
- **Looking for interesting science proposals**
  - Codes that scale to the tens of thousands of tasks
  - Identify applications and personnel (1/2 FTE) by 10/1/03
  - Project plan by 10/30/03
  - Contact: Don Dossa dossa1@llnl.gov
- **Schedule**
  - Science runs – maybe 1Q 2005
  - BGL simulator available soon on ALC

Blue Gene/L

	ASCI White	ASCI Q	Earth Simulator	ASCI Purple	BlueGene/L†
Machine Peak Speed (Tflop/s)	12.3	20	40	100	180 / 360*
Total Memory (Tbytes)	8	22	10	50	16–32
Footprint (ft.^{2})	10,000	20,000	34,000	12,000	2,500
Total Power (MW)	1.0	3.8	10	4.5	1.2
Cost (M$)	~100	~200	~350	~250	<< 100
Installation Date	9/2000	~9/2002	2/2002	12/2004	~12/2004
No. of Nodes	512	4,096	640	197	65,536
CPUs per Node	16	4	8	64	2
Clock Frequency (MHz)	375	1,000	500	~2,000	700
Power Dissipation/Node (W)	2,000	920	16,000	23,000	15
Peak Speed/Node (Gflop/s)	24.0	7.3	64.0	512	2.8
Memory/Node (GiB)	16	8	16	250	0.25–0.5
Memory Bandwidth (TB/s)	8	19	160	130	360
Memory Latency (cycles)	140	330	–	280	70
MPI Latency (μs)	25	4.5	6–20	5-10	7
Interconnect Bandwidth (B:F)	0.042	0.085	0.13	0.13	0.75
Bi-Section Bandwidth (B:F)	0.04	0.04	0.03	0.06	0.008

† target specifications

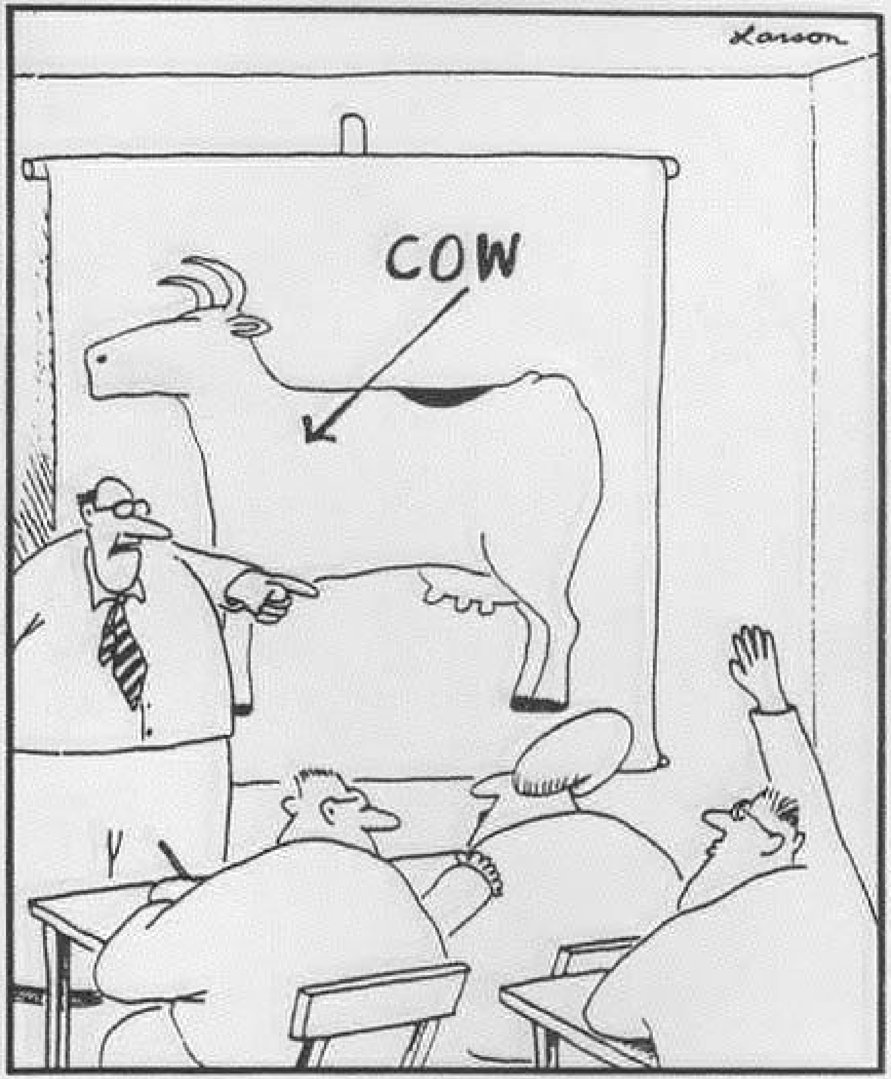
* comm. co-processor mode / symmetric mode

Terascale Simulation Facility

- **Home for next-generation computer systems**
    - Designed to accommodate two 100+ teraOPs systems
    - First system will be the 60+ teraOPs ASCI Purple system in 2004
    - Second system will be Blue Gene/L
    - Office space for 288 scientists, engineers and support staff
    - Groundbreaking: 4/4/02
    - Completion: 2006

- **A few metrics**
    - 253,000 ft^2 total
    - 2 machine rooms totaling 48,000 ft^2
    - 22 MW power total
    - 9 MW for machines
    - $92 million cost

Terascale Simulation Facility



"Yes … I believe there's a question there in the back."